Random Walk with Wait and Restart on Document Co-citation Network for Similar Document Search

Masaki Eto
Gakushuin Women's College
Tokyo, Japan
masaki.eto@gakushuin.ac.jp

ABSTRACT
One of the latest algorithms for computing similarities between nodes in a graph is Random Walk with Restart (RWR). However, on a document co-citation network for similar document search, computing transition probabilities remains difficult. To solve the problem, this paper proposes a Random Walk with Wait and Restart (RWWR) algorithm, which contains a new technique for adjusting the transition probability by incorporating a "self-returning" edge into the normalization. To evaluate its effectiveness empirically, the search performance of two retrieval methods using RWWR was compared to a method using the standard RWR; the performance was measured by average precision and nDCG. The experiment was conducted on a test collection created from the Open Access Subset of PubMed Central, and the results indicated that the RWWR methods tend to outperform the standard RWR method.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Retrieval models

Keywords
Citation searching, Similar document search, Random walk

1. INTRODUCTION
One of the latest algorithms for computing similarities between nodes in a graph is Random Walk with Restart (RWR). The RWR algorithm iteratively investigates the entire network to calculate the similarity between a seed node and each node in a network. Specifically, the walker starts at a seed node, then either proceeds to the connected node based on a transition probability calculated by edge weights, or returns to the seed node. The vector for the probability that the walker stays at individual nodes is defined as:

$$p = (1-r) \times T \times p + r \times s \quad (1)$$

where p is an n-dimensional vector (n is the number of nodes in a network), r is a return probability, T is a transition probability matrix, and s is an n-dimensional vector with 1 for the seed node and 0 for the rest. This equation is applied recursively until convergence, and then each vector value of p is used as a degree of similarity to the seed.

This method can be applied to a document network by citation linkage, which is a kind of directed edge, for implementing recommender systems (e.g., [2], [4]). In such systems, the search query is a seed document known to be relevant to the information needs of a user.

Copyright is held by the author/owner(s).
RecSys 2014 Poster Proceedings, October 6-10, 2014, Foster City, Silicon Valley, USA.

Another type of document network is the co-citation network (see Fig. 1), which is often used in the field of scientometrics (e.g., [3]). This network is composed of document nodes connected by co-citation linkages, which are undirected and each linkage indicates a relationship between a pair of documents concurrently cited by a third document. In addition, each edge has a weight, i.e., strength of the relationship based on the number of documents citing both nodes; e.g., five documents co-cite Document A and C_1 in Fig. 1 (citing documents are not shown).

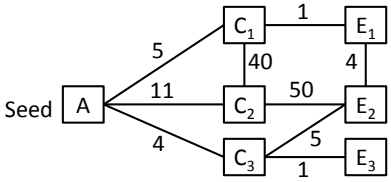

Figure 1. A co-citation network.

This paper explores a suitable technique of applying RWR to a document co-citation network for similar document search. Although RWR works well on a document co-citation network for it [1], a problem remains in computing the transition probability. That is, the standard RWR may unreasonably calculate a transition probability for an edge from a current node to the next node partly because the edge weight is normalized according to the sum of weights of edges connecting to the current node, often unexpectedly causing a weak edge to have a higher transition probability than strong edges. In Fig. 1, the transition probability "E_1 to C_1" including only one co-citing document is higher than "C_2 to A" obtained from 11 co-citing documents, i.e., "E_1 to C_1" is 0.200 as 1 / (1 + 4) and "C_2 to A" is 0.109 as 11 / (11 + 40 + 50).

Clearly, in the case of a co-citation network, an alternative normalization approach for calculating T of Eq. (1) is needed to keep a high transition probability for an edge constructed from heavily co-cited documents such as "C_2 to A." This paper proposes a Random Walk with Wait and Restart (RWWR) algorithm, which contains a new technique for adjusting the transition probability by incorporating a "self-returning" edge into the normalization (Fig. 2).

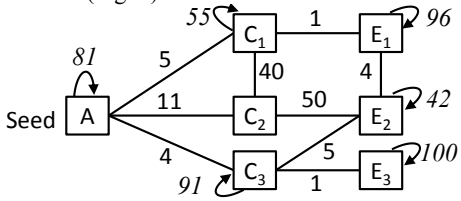

Figure 2. A co-citation network with self-returning edges.

2. PROPOSED TECHNIQUE
The self-returning edge is directed from a node to itself, and keeps the walker staying at the current node for a while. If the sum of

weights of undirected edges connected to node *v*, denoted by *outlink(v)*, is small, then it can be reasonably assumed that the walker is reluctant to move from *v* because of weak edges to other nodes. This paper defines a more appropriate transition probability by adding a weight of the self-returning edge at node *v*, denoted by *w(v)*, to the denominator of normalization.

In order to estimate *w(v)*, the maximum value of *outlink(v)* in the network, denoted by *max_outlink*, can be used:

$$\max\_outlink = \max_{v \in V} outlink(v) \quad (2)$$

where *V* is a set of all nodes in the network. This paper explores two methods for estimating *w(v)* based on *max_outlink* as follows. Method 1 uses the difference between *max_outlink* and *outlink(v)* as the value of *w(v)*:

$$w(v) = \max\_outlink - outlink(v) \quad (3)$$

In Fig. 2, because *max_outlink* is 101 given by outlink(C_2), w(A) becomes 81 (= 101 − (5 + 11 + 4)), and therefore the transition probability of "A to C_1" is 0.05 by using the sum of *w(v)* and *outlink(v)* as 5 / ((5 + 11 + 4) + 81).

Method 2 aims to avoid w*(v)* from becoming too large. If w*(v)* is too large, then the transition probability becomes too small and therefore the walker does not move around on a network adequately; for example, in Method 1, "E_3 to C_3" is 0.01. Hence, in Method 2, w*(v)* is adjusted by using the value of *outlink(v)* as the upper limit, which means that the upper limit of a transition probability is 0.500. Specifically, Method 2 calculates w*(v)* as:

$$\mathrm{w(v)} = outlink(v) \times d(v) \quad (4)$$

where *d(v)* shows the degree of difference between *max_outlink* and *outlink(v)*; *d(v)* ranges from 0 to 1 and is calculated by

$$d(v) = \frac{\max\_\mathrm{outlink} - outlink(v)}{\max\_\mathrm{outlink} - \min\_\mathrm{outlink}} \quad (5)$$

where *min_outlink* is the minimum value of *outlink(v)* in a network. In Fig. 2 where *min_outlink* is 1 given by outlink(E_3), w(A) becomes 16.2 as (5 + 11 + 4) × ((101 – (5 + 11 + 4)) / (101 − 1)), and therefore transition probability "A to C_1" is 0.138 as 5 / ((5 + 11 + 4) + 16.2).

3. EXPERIMENTS

To evaluate the effectiveness of the proposed method empirically, the search performance of two retrieval methods using RWWR was compared to a method using the standard RWR (Baseline), and the performance was measured by average precision (AP) and normalized Discounted Cumulative Gain (nDCG).

To create a test collection, about 152,000 documents were selected from the Open Access Subset of PubMed Central, under the condition that each document had at least one citation linkage with another document in the subset.

In the experiment, it was assumed that a seed document was given by a user as a search query. The test collection contained 100 seed documents selected randomly from all documents under two conditions. First, documents cited by 10 or more other documents were extracted. This condition yields co-citation networks with sufficient numbers of documents. Second, each co-citation network contained one or more relevant documents. By using 100 seed documents, 100 co-citation networks were respectively created from documents within two hops from each seed.

In the experiments, whether a document was relevant was determined by the degree to which it shared MeSH Descriptors with the target seed document. Specifically, the Jaccard coefficient (JC) was used; when AP was calculated, documents whose JC was 0.2 or more were regarded as relevant, and nDCG used a relevance score of 3 for documents whose JC was 0.4 or more, 2 for documents whose JC was 0.4–0.2, and 1 for documents whose JC was 0.2–0.1.

Search runs for 100 seed documents were executed by each method and then scores of AP and nDCG per seed document were measured. In the ranking process, when two or more documents had the same score, their ranks were randomly assigned for tie-breaking. The results are shown in Fig. 3 and Table 1 below.

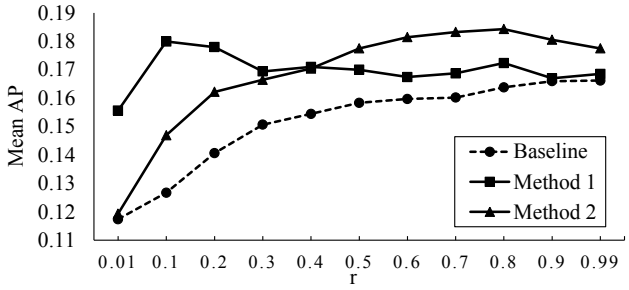

Figure 3. Mean scores of AP.

Mean scores of AP are shown in Fig. 3, where the horizontal axis indicates the value of *r* of Eq. (1).

Table 1. Comparison baselines with the proposed methods.

	Baseline	Method 1	Method 2
AP	0.166 (r = 0.99)	0.180 (r = 0.1)	0.184* (r = 0.8)
nDCG	0.640 (r = 0.9)	0.650 (r = 0.1)	0.651** (r = 0.8)

* P < 0.05, ** P < 0.01

Table 1 compares the best scores of the three methods. As shown, both of the proposed methods outperformed the baseline. In addition, the paired t-test indicated a statistically significant difference for Method 2.

4. CONCLUSIONS AND FUTURE WORK

This paper proposed a Random Walk with Wait and Restart (RWWR) algorithm on a document co-citation network for similar document search. The experiment results indicated that the RWWR method tends to outperform the standard RWR method. The method will be applied to a larger network in a future study.

5. ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number 23700289.

6. REFERENCES

[1] Eto, M. 2013. Document retrieval method using random walk with restart on co-citation network, Workshop on Informetric and Scientometric Research (METRICS 2013).

[2] Gori, M. and Pucci, A. 2006. Research paper recommender systems: A random-walk based approach. In *Proceedings of IEEE/WIC/ACM Web Intelligence*, 778-781.

[3] Hu, Y., Sun. J., Li, W., and Pan, Y. 2014. A scientometric study of global electric vehicle research. *Scientometrics*, 98, 2, 1269-1282.

[4] Küçüktunç, O., Saule, E., Kaya, K., and Çatalyürek, Ü. 2012. Direction awareness in citation recommendation, In *Proceedings of the 6th International Workshop on Ranking in Databases* (DBRank'12).